BOG 8: System Management, Administration, & Job Scheduling

ASCR Workshop on Extreme Heterogeneity in HPC
23-25 Jan 2018

BOG 8 Contributors

Moderator(s): Rebecca Hartman-Baker & Paul Peltz

BOGists: Ray Bair, Bob Colwell, Jeanine Cook, Tina Declerck, Doug Jacobsen, Zhiling Lan, Vitus Leung, Barry Rountree, Rob Ross

BOG 8 Capability Targets for Extreme Heterogeneity

BOG 8 brainstorming and discussion of capabilities that will be needed in the 2025-2035 timeframe to make increasingly heterogeneous hardware technologies useful and productive for science applications. The purpose of this workshop is to target the 2025-2035 timeframe for machine deployments. If ECP already has a plan to deliver on a particular piece of technology then it is out of scope of this workshop. We are charged to look beyond ECP and to help predict the future technology that will need additional funding to deliver. What are the challenges that we think need to be addressed to support and deploy heterogeneous hardware technologies?

Notes: https://docs.google.com/document/d/1mHyvVJU8VRspzREHwm0-2Qnx5uE2O6bHBYYEg6M_vRo/edit?usp=sharing

BOG 8 Targets for 2030

Target 1: Systems Management Requirements

- Managing diverse array of components within a HPC system
  - Varying architectures for compute nodes
  - Varying support service nodes
  - Ability to manage different EH components that are procured/decommissioned throughout the lifetime of other HPC and EH components
- HPC Management Software
  - Orchestration state engine to manage data center HPC resources
  - Always available computing (external data ingestion)
    - Automated detection of failures and response
  - Integrated Configuration Management Software
    - Standards to manage EH components and subscribe to our configurations
- HPC Provisioning
  - More robust tools for management of the complex software ecosystem necessary
  - Custom binaries and vendor specific implementations will not support the diversity of needs of EH systems that may encompass multiple HPC resources (Viz, Instruments, Data movement)
- Security
  - How do we trust and verify the data we are ingesting from a variety of IoT, instruments, etc.

BOG 8 Targets for 2030

Target 2: Scheduling and Workflows

- Power Aware Scheduling
  - Job scheduling based on an entire facilities power consumption/limit/rates
  - Power ramp up/down capabilities
  - Dynamic power management of resources, in job/workflow - power donation, request power
  - ML of application power profiles
- HPC resource scheduling
  - User requestable features for bandwidth, storage capacity, burst buffers, etc.
  - Managing diverse workflows, pre/post-staging of data to or from either external or internal data storage facilities or instruments (post SKA/LHC)
  - Increasing utilization of EH components
    - How much of the system utilization that we report is fully utilized by the jobs?
    - backfilling based on node utilization, i.e. job oversubscription
- Scheduling API
  - Communication across EH objects, facilities, external instruments

BOG 8 Targets for 2030

Target 3: Diagnostic Tools and Monitoring

- Failure Management/Recovery
  - Extreme heterogeneity increases the complexity and probability of failures
    - Request failure scenarios for ML from vendors to help diagnose problems
    - Public ML training databases for EH components
  - Better monitoring of components necessary
  - Functional and performance tests required to return failed components to service
  - Job recovery mechanisms during event failures, restarting workflows from point of failure
- Monitoring and Log Analysis
  - Aggregating and analyzing logs from a variety of components is a challenge
  - Disparate logging creates challenges to correlate job failures to component failures
  - Applying root-cause modeling to analyze failures

BOG 8 Current Capability

Capability 1: Everything is its own source of truth with no communication between the components. This applies to monitoring, logging, system management, etc.

Capability 2: Power management limited to x86 and no insight into accelerators.

Capability 3: Burst-buffer co-scheduling implemented with Greedy algorithm, limited true multi-point optimization of scheduling resources.

BOG 8 Challenge Assessment

Discussion to identify research required to get from where the capabilities are now to where to where they need to be by 2030.

Orchestration agent that handles all requests of not just the system but the facility.

All we have is x86 power control and more is needed in co-design efforts with accelerator vendors so we understand what they are giving us and they understand what we need. Need more research into what we need with regard to power control for accelerators and other resources.

BOG 8 Possible Research Directions Summary

- PRD 8.1 - Orchestration agent for data center awareness and state
  - EH, facility power, and external components can publish and subscribe to this agent
  - Vendor agnostic APIs and data schemas for systems management and tools
- PRD 8.2 - Scheduler Improvements
  - Multi-point optimizations for co-scheduling heterogeneous elements (attempting to saturate all compute elements) or facility resources (like power or network)
  - Many system federation to enable workflows informed by remote data or remote capabilities

BOG 8 Possible Research Directions Summary

- PRD 8.3 - Security
  - How do we trust and verify the data we are ingesting from a variety of IoT, instruments, etc.
  - Securing the data supply line into the EH systems from uncertain origins
  - Multiple users running on the same node sharing EH components presents security concerns
    - Isolating user applications to zones of trust
      - Code trust: long standing vs new/unknown code
    - Trusted executable based on static analysis or sandbox runtime
- PRD 8.4 - Monitoring and Log Analysis
  - Use ML to learn failure scenarios and report predicted failures to signal jobs, and state management system
  - Machine Learning databases provided by vendors and community contributed repos
  - EH component monitoring
    - Detect failures, or degraded performance
    - Functional and performance unit testing to return components to service

An Integrated View of the PRDs

<table>
<tr><th colspan="2">Stacked</th><th>Cross-Cutting</th></tr>
<tr><td colspan="2">PRD 8.1: Orchestration agent for data center awareness and state</td><td rowspan="2">PRD 8.3: Security
● BOG 2: Data Management
● BOG 4: OS/Resource Management</td></tr>
<tr><td>PRD 8.2: Scheduler Improvements</td><td>PRD 8.4: Monitoring and Log Analysis</td></tr>
</table>

PRD 8.1: Orchestration agent for data center awareness and state

System Management becomes more complex when the number of different elements that need to be configured, scheduled, monitored, and run efficiently increases. The potential for cost decreases on some components also increases the possibility of needing to add new resources while maintaining utilization. A mechanism such as an orchestration agent for managing all aspects of the system and the facility is needed to both standardize and simplify management tasks.

- Research Challenges:
  - An agent needs to be able to obtain data in a standard format. Research is needed to understand and develop common APIs that define the various aspects of system management including (but not limited to):
    - The state of the system for monitoring,
    - A way to allow vendors to define configuration requirements for their resources that can be automatically included in the system configuration
    - Required components for scheduling
  - Research into how to tie facility related information to the agent could inform best use of resources such as power. This information would also need to be shared with the workload manager in order for job scheduling to reflect the requirements. This requires not only APIs for getting access to the information but a mechanism for setting limits that is flexible enough for different site uses.
- Research Approach:
  - This requires reviewing system and facility related configuration, monitoring, and workload elements at various facilities to be able to define a rich set of APIs that can handle both existing and unknown new technologies and provide an automated interface to integrate these into the agent. An additional element is how to best maintain the information for rapid access

PRD 8.2: Scheduler Improvements

Extreme Heterogeneity will provide users with a multitude of choices on how to manage and optimize their applications to utilize different components of the system. In order to efficiently utilize a possible variety of components within a node, it will be desirable to go to a multi-tenant scheduling model to fully utilize a node's available resources. In order to better utilize EH resources it will be necessary to have multi-point optimizations for co-scheduling of these EH resources. This will allow HPC facilities to better saturate all compute elements or facility resources such as power or network.

- Research Challenges
  - Federating scheduling across all of the various elements in a data center
  - Multi-tenant usage of nodes and effective utilization of all compute resources
- Potential research approaches and research directions
  - Multi-resource co-scheduling and future planning to maximize utilization of disparate resources
  - Utilizing performance counters to determine node resource usage
  - Researching possibilities for data center coordination through the state engine in PRD 8.1
- How and when will success impact technology?
  - Multi-tenant node usage will increase utilization and the scientific cycles the systems are capable of delivering
  - Scarce resources will be better utilized because jobs and workflows will be better coordinated through the scheduler

PRD 8.3: Security

Extreme Heterogeneity will provide users with a multitude of choices on how to manage and optimize their applications to utilize different components of the system. In order to efficiently utilize a possible variety of components within a node, it will be desirable to go to a multi-tenant scheduling model to fully utilize a node's available resources. Security is a concern due to this, and research is necessary in order to secure zones of trust on a node. The multi-tenant security concerns are compounded by instrument data that could be ingested into the system from unknown or compromised sources. Research will also be necessary in securing the data pipeline from instrument to system.

- Research Challenges
    - Complete isolation of a user's namespace on shared resources, e.g. [NV,D,MCD]RAM
    - Static analysis or sandboxing of a user's application in order to ensure security
    - Ensuring data integrity/authenticity from creation to consumption
    - "Beyond POSIX" standards to implement security
- Potential research approaches and research directions
    - Kernel per user that only has accesses the components of the node requested by the scheduler
    - Utilization of a technology like blockchain to ensure authenticity of data
- How and when will success impact technology?
    - Multi-tenant node usage will increase utilization and the scientific cycles the systems are capable of delivering
    - Securing of data and isolation of namespaces reduces the attack surface exposed by EH systems

PRD 8.4: Monitoring and Log analysis

Monitoring and logging of facility resources is and continues to be a challenge. This challenge will be even more problematic when extreme heterogeneity becomes commonplace. The monitoring of a variety of EH components to ensure they are available and performant will be a necessity in order to rapidly offline or fix components. The prediction of component failures by using machine learning to signal jobs that a particular component will likely fail which would allow the users to respond to that failure scenario and migrate off of that failing resource.

- Research Challenges
  - Machine Learning needs structured data. Logging standards would need to be created for vendors to implement.
  - Fast testing of components and job/workflow component testing to ensure the EH components of the job are healthy
- Potential research approaches and research directions
  - Research into job migration from failed/failing components
  - Curation and development of a machine learning database/repository for failure data
- How and when will success impact technology?
  - Increased system availability metrics to make EH components more highly available
  - Prevent scientists from having to redo lost work due to failures and waste computational time and resources